An Arizona Model for Web Preservation and Access

Richard Pearce-Moses
Presented to the ALA Government Documents Roundtable (GODORT)
June 2005

Draft v3a

The Arizona State Library is the oldest cultural institution in the state. Since it was founded with the territory in 1863, it has been mandated to collect and provide access to official reports and publications for current and future use.

The World Has Changed

I don't want to state the obvious, but a lot has changed since then. I think the biggest change has been the rise of the web. Vannevar Bush's vision of memex was realized at last.[1] I don't know that Bush realized that this system – at least as realized – would not only make the universe of information more readily accessible, but would dramatically expand the size of that universe. On average, state agencies' Web sites contain more than 300,000 documents at any given time. I am concerned that many – maybe most – librarians and archivists don't fully realize that the change has not been evolutionary but revolutionary.

Part of the problem is that the use of 'publication' as a criteria for our collections no longer works in the age of the web. Before the web, 'publication' carried the connotation of something printed in large quantities for public distribution. The costs of printing limited the number of publications. But low costs of distributing materials on the web means that much more is published.

The increased number of documents on the Web promises a vastly richer collection of publicly available reports and publications. Now, it's much easier to locate and capture fugitive documents. These are very good things for depository programs.

Libraries Must Respond

However, Web documents present a number of challenges to traditional ways of curating a print-based collection.

1. We must rethink the scope of our collections; we can no longer use 'published' as the primary criteria to distinguish what we collect for the depository collection. The web is used to distribute ephemeral documents, in addition to official reports and publications, blurring the distinction between which documents should be added to the depository program and those with limited value.

2. We must find ways to scale our curatorial practices to the enormous number of web documents and the need for an innovation and adaptation of traditional workflows and practices. Web documents often lack the formal elements of printed reports and publications; without a cover sheet or title page, finding the information necessary to describe the documents can be a challenge. Where printed documents have a simple and familiar structure – ink on paper sheets with a binding that defines the content's sequence and boundaries – Web documents are often created using specialized software and may contain links that blur the document's boundaries.

3. In particular, we must understand the profound impact on core skills and preservation programs.

[1] See "As We May Think," *The Atlantic Monthly* July 1945. An online version is available at http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm.

Curation

Today, I want to focus on the problem of curation. I use the word 'curate' to refer to a wide range of activities that all cultural professions share: identification and selection, acquisition, description, reference, and preservation. When I speak of curating a collection, I'm talking about the principles underlying those practices, as well as the policies, procedures, and workflows that are often so familiar that we take them for granted.

As a profession, we must find new ways to curate our collections in the digital era. I firmly believe that *what* we do will remain fundamentally the same, but *how* we do those things in a digital environment will change significantly. I believe we must reengineer our profession. This is not merely a matter of using technology to help us do the same thing faster or cheaper. It means rethinking the purpose and the workflow of what we do.

Curatorial Responses

In looking at how others struggle with curating collections of web publications, I see two general approaches.

> First, a 'bibliocentric' model based on a traditional library processes of selecting documents one by one, identifying appropriate documents for acquisition; electronically downloading the document to a server or printing it to paper; then cataloging, processing, and distributing it like any other paper publication. This approach can capture a low volume of high quality content. However, it cannot be scaled to the massive numbers of Web publications without a large increase in human resources.
>
> Second, a 'technocentric' model focused on software applications designed to capture everything. This approach trades human selection of significant documents for the hope that full-text indexing and search engines would be able to find documents of lasting value among the clutter of other, ephemeral content. This approach essentially transfers the work of selection from the libraries to the patron.

Arizona Model

THE ARIZONA MODEL

In Arizona, we began rethinking basic assumptions about how we were to integrate web documents into our state depository program. We began developing a new model – what is now referred to as the Arizona Model – for our workflows using a cross-disciplinary approach that combined the strengths of library, archives, and information technology.

The Arizona Model reflects a fundamental shift from a bibliographic approach to an archival approach for curating collections. We believe that an archival approach makes sense because websites are very similar to archival collections. A website is a collection of documents with common provenance; the documents were created or received in the course of business. A website is – in most cases – organized into directories that group documents by subject or function, in the same way that an archival collection is organized into series of documents organized by subject or function. Finally, websites have far too many documents for us to be able to curate at the item level; like archival collections, we must first curate these materials as aggregates and reserve item-level control for only the most important documents.

The Arizona Model articulates ways we might change the curatorial practices of our state documents depository, and ultimately how we might curate all our collections at the Arizona State Library and Archives. The model addresses identification, appraisal and selection, description, reference, acquisition, and access. It reflects our exploration of how the principles and goals of those activities, as well as the practical activities to accomplish them in the digital era. I use 'might' because the model should be understood as a hypothesis that must refined as we test it through practical application.

IDENTIFICATION

The first challenge we faced was identifying which of the millions (maybe billions) of documents on the web were in-scope. Here's where the archival approach comes in. Rather than looking at the documents, we look at the provenance. We look for just those websites that operated by state agencies. Even so, how will we find those websites among the 64 million sites on the web?[2]

You might think someone has a list. In fact, Arizona state government, like many in the West, is highly decentralized and each agency is largely autonomous. In many ways, that independence reflects both the spirit of the West and the spirit of the Web. That spirit also makes the Library and Archives' job harder. We had to build the list.

We started with the hypothesis that any state website would be referenced on at least one other agency website. So, we began building a list of all websites referenced on state agency websites. We used spider software to look at a few key state websites that were rich with links to state agencies, such as the state portal and the governor's website. The spider's initial run returned a list of some 50,000 links, way too many documents for us to analyze. We took that report and extracted a list of about 1,500 distinct domains – everything up to the first slash in a URL. That was a number that we could work with.

Looking at the list we could easily exclude or include entire websites based on their provenance. It was immediately obvious that adobe.com was out of scope and that www.azgovernor.gov was in scope. In fact, we needed to take time to look at about 500 sites. And we found lots of websites that we were interested in. The domain www.phoenixvis.net gives no clue that the site is operated by the Department of Environmental Quality. At the same time, the domain www.az51.com – based on a freeway – sounded like it might be operated by the Department of Transportation. Transportation disavowed responsibility. But here it gets interesting. Transportation requires contractors to maintain websites about road projects to keep the public informed. While Transportation did not view these as state documents, the Library and Archives does because those documents were created using state funds in the course of performing state business.

Evaluating the list took less than a week's time. We intend to repeat this process on a regular basis (monthly or quarterly), and we expect the evaluation to take significantly less time on such sequent runs. The database will remember our decision that adobe.com is out of scope, so we don't need to reevaluate it.

Once we identified in-scope websites, we began building a database to associate the website with its creator. The database is, in effect, a very powerful authority file. In addition to tracking variant forms of an creator's name, it also provides us with a place to store its administrative history. We use the database relate website creators to each other; for example, we can link predecessors and successors, as well as parent-child relationships. Ultimately, I envision this database as a taxonomy of state government.

The database also stores access points to help discover agencies by "subject" using a controlled vocabulary. I'll come back to how I see this creator-level metadata being used when I talk about description and access.

SELECTION/APPRAISAL

We've already made an initial, coarse appraisal decision by determining if a website is in-scope. This process is another example of the archival approach, which looks at aggregates rather than individual documents. When we say 'no', we may be making decisions about thousands of documents.

We take the aggregate approach a step further, though, when we look at how the creator has organized the documents. In a traditional paper environment, records creators organize documents by content and

[2] "Jakob Nielsen's Alertbox, June 1, 2005: Alertbox: Ten Years" (http://www.useit.com/alertbox/20050601.html). Checked 2 June 2005.

activity; purchase orders, personnel files, and program records are kept in distinct series. (There are always exceptions to this rule; it's not unheard of for an office to keep everything in a single, alphabetical file series.) Archivists typically do not select individual documents, but select the series as a whole. As a rule of thumb, purchase orders do not have permanent value so the series is discarded; there might be one or two of long-term interest, but archivists are willing to let those few go rather than trying to weed through them all. Contrawise, a series of executive correspondence may be acquired in its entirety, even though it may contain some letters of no value.

Web masters also tend to organize their documents by subject or activity, so we can apply the same approach to web documents.

What we need is the ability to see the organization of the website. The problem is that web masters aren't likely to give us remote access to their systems; a reasonable decision given security considerations.

The list of URLs we created to search for in-scope websites can be used to reveal that structure. We analyzed all the links for an in-scope domain to build a hierarchical representation of the directories based on the subdirectories in the URLs. For example, the Department of Water Resources[3] has the following top-level directories, and looking at the files in those directories we can rapidly make some selection decisions. Again, we're making decisions about groups of files, not individual files.

_derived/	X	System files and program code.
Browse/	√	While the name of the directory give no clue, a sample of the directories and files contained under Browse make it valuable as a whole
ContactUs/	X	These documents are of ephemeral value.
Employment/	X	These documents are of ephemeral value.
FAQ/	X	These documents are of ephemeral value.
Forms/	X	We don't collect blank forms.
InfoCentral/	X	A document imaging system that is inaccessible to web spider software. We work directly with the agency to get these items.
LectureSeries/	√	The contents are valuable, but we also noted that nothing has been added in a long time. We'll capture it, but it's not likely to grow.
News/	√	Although the Library does not generally capture press releases, the Archives does.
Publications/	X	This directory name is misleading. It sounds like exactly what we want, but it contains only pages that link to things that are under other (captured) directories.

The Browse directory is subdivided into a number of subdirectories, and that subdirectory is the rich vein of documents we want to mine.

Conservation/	√	Specific documents targeted towards consumers, such as gray water, xeriscaping, and low-flush toilets.
Drought Task Force	√	Policy and planning documents from the Governor's office.
Management Plans	√	Policy and planning documents for water use in metro areas.
Surface Water	√	Agreements and information about the use of surface water.
Water Quality Fund	√	Publications about financial investments to protect water.

As you might expect, in Arizona the Department of Water Resources is important and has a large website with around 5,000 documents in about 50 directories and subdirectories. If we spent an average of five

[3] The series and descriptions of documents in these series are loosely based on the Department of Water Resources' website, but the examples here were simplified for illustrative purposes. In fact, the site is more complex than represented here. However, much of that complexity comes from directories that can quickly be deselected because they contain software and support files rather than documents for acquisition.

minutes selecting the documents individually, it would take three solid months – with no time spent on anything else. If we spent five minutes on the 50 directories and subdirectories, we can be finished in less than a day. I believe a day is a realistic amount of time to spend on a relatively complex website. Some directories will clearly take more than five minutes, but many take less.

This general description does not account for a number of variations. Some agencies have publications scattered across several web servers that use different domains, and the model has to be adjusted to see these several servers as an integrated website. Several websites may share a single domain; for example, different divisions within an agency may have websites that are, in effect, directories off the agency's website (lib.az.us/archives/, lib.az.us/btbl/, lib.az.us/ldd/, lib.az.us/museum/). However, the approach here is easily adapted to these variations.

DESCRIPTION

Once we know those series that we want to acquire, we will need to describe them. Again, the archival approach is to – in general – work with aggregates, not individual documents. Archival finding aids are commonly formatted like an outline, with the series (directories and subdirectories) are major headings with lists of folders under each heading.

Here is a clear different between bibliographic and archival description. Cutter's objective for a dictionary catalog was to show the patron what the library held. If a work wasn't in the catalog, it wasn't in the library, and the patron need look no further. Archivists show patrons the organization of a collection and the most likely places to look, so that a patron has to look at a small number of items. There is no assurance that the archives holds relevant materials.

Given the enormous number of documents on the web, I can't imagine that we will ever be able to describe them all using a traditional bibliographic approach. Even using an archival approach to eliminate the irrelevant materials, I can't imagine that we'll have the time to catalog them all. At the same time, because the documents are in electronic format, I think we can use machines to provide greater detail than a traditional archival finding aid.

Taking an archival approach, we describe series, not individual documents. Here again, the difference between bibliographic and archival description is significant. Where bibliographic catalogers transcribe, archival catalogers must supply. We are using humans to describe the abstract and computers to describe the concrete. Collections and series are, in many ways like forests; they are abstract concepts. You can't touch a forest. You can only touch the trees in the forest. Human describe the forest (the collections and series), and computers list the trees (the documents in the forest).

The first step is to establish the title for the series by translating the directory name into natural language. Web masters want to keep URLs short, so they will use abbreviations, acronyms, or other devices to represent the name of the series. On Water Resources' website, the directory GDTF needs to be expanded to Governor's Drought Task Force, and WQARF needs to be expanded to Water Quality Assurance Rotating Fund.

The next step is to describe the contents of the document in general terms by writing a scope note. The principal information to record is the reasons the Library and Archives selected these materials – why are they important? why do they merit preservation? The scope and contents note is not intended to be a scholarly dissertation, but to provide users with an introduction to the materials and to record the cataloger's knowledge of those materials.

Finally, we assign descriptive metadata using a controlled vocabulary. Because we're working with aggregates, we assign metadata that is true for all the documents in that series.

name="Creator"	Governor's Drought Task Force	Rural Watershed Alliance
name="Subject"	reservoirs	ground water

name="Subject"	drought	water conservation
name="Subject"	potable water	agriculture
name="Type"	planning	reports

The Library and Archives is using the Jessica Tree, a thesaurus designed for high-level analysis of government information developed by a number of state libraries. However, other vocabularies could be used.

Once the documents are harvested, we will be able to use the metadata to produce a finding aid that organizes the documents by collection, series, and subseries. Listing individual documents in a finding aid is generally impractical because of the time involved. Because the documents are in electronic format, we can potentially harness the power of the computer to generate this detailed list of titles.

ACQUISITION

Normally libraries and archives describe materials after they acquire them. In our approach, acquisition is the final step and will be fully automated. The software will look at the databases we've built to determine which websites and which series on those website contain documents we want to harvest. It then begins the process of pulling all those documents off the web. As it does so, it packages the document with metadata based on what we've recorded in the databases. Creator metadata comes from the site properties database. Subject headings (using a controlled vocabulary) can be taken from the site analysis database. It will also capture administrative and preservation metadata, such as date harvested and an MD5 hash value to enable us to demonstrate the content has not changed over time.

The packaging tool also looks for any user-supplied metadata within a document, such as title or creator. We'll be investigating the limits of user-supplied metadata. First, we recognize that only a small percentage of documents include user-supplied metadata. Second, we also know that some sites built using templates use the same title for many documents, and in a few instances the title is a default value along the lines of "Insert title here." I mention this to illustrate a fundamental principle of the project: we're not going to let the perfect get in the way of the possible.

In addition to these access points that used controlled vocabulary, we are investigating the use of other tools to automatically supply additional metadata. For example, when we spot a series that assigns a generic title to all documents, we may direct the harvester to find a noun-phrase in the document that could be used to distinguish it from others. We're also interested in exploring a number of tools that could autogenerate a scope note or classifications. I have a healthy skepticism about computers' ability to do this sort of analytical work, but we want to look at how these tools can help improve access. Remembering our philosophy of the perfect not being a barrier to the possible, we are not looking for AI tools that work perfectly, but for those that improve access significantly.

REFERENCE

The final function I'll talk about today is reference. We must consider how we will help our patrons find these documents.

The simple answer is full-text searching. Google has set the standard for user expectations. At the same time, as librarians and archivists, we know that Google's algorithm works less well in these sort of closed, institutional collections. And, I think we can do better by taking advantage of the metadata we assigned at the creator and series level.

For me, the problem with Google (and most other search engines) is that they returns a very long list of documents. The good news is that the documents at the top of the list are often sufficient. The bad news is that sometimes the needle you're looking for is buried in the haystack. For example, I was researching

a moving image production technique that uses a B-reel. I Googled 'B-reel'; when I got the results, I immediately realized I was out of luck. B-reel is synonymous with 'blue movie'. Now, you may find this shocking, but there's lots of information about blue movies on the web. So much, in fact, that I never could find any documents about the camera technique. If the results could have been categorized into groups based on content, I could have found what I would looking for.

Your search for: water, Phoenix

Found documents in the following categories

water (500+)	water conservation (357)	Salt River Project (210)
drought (110)	flood control (98)	xeriscape (25)

Found documents from the following agencies

Water Resources (135)	Governor's Drought Task Force (102)	Phoenix (87)
Maricopa County (84)	Corporation Commission (35)	

The challenge will be to find search-engine software that can take advantage of the metadata to organize masses of data into results that will be more useful than a ranked list. I believe such a tool is particularly valuable because the categories can help patrons refine a vague, ill-formed query by suggesting more specific topics. We have identified a number of products that may be able to do this, if we can figure out how to pay for them.

PRESERVATION

The one activity I will not talk about is preservation, but I will make one observation. Preservation has two meanings. First, conservation. Second, pear preserves. We do not yet have the tools or resources to ensure long preservation. What we are attempting is to preserve documents in the second sense; we want to capture information that's on the web now so that when the long-term tools are available we will have something to save.

THE WEB ARCHIVES WORKBENCH

In the midst of thinking about an archival approach, OCLC and UIUC approached us about partnering with them on a research project they were proposing to LC's NDIIPP. The project would, among other things, bring in four other state libraries to help develop and test the archival approach, and it would build software tools to help implement that approach. I'm happy to tell you that LC awarded UIUC and OCLC the grant.

Initially the workbench will have four tools that work together. We used the workbench metaphor to suggest that a curator may choose to use different or additional tools in a coordinated fashion.

1. Domain Tool: The first tool is designed to help us identify in-scope websites. It spiders sites, harvesting and analyzing links to build a list of distinct domains. It alerts us when new sites appear or known sites disappear. It keeps a record of our decisions as to whether sites are in-scope so that we don't repeat that work.

2. Properties Tool: The second tool is the database of creators responsible for the websites. It includes the rich authority information about the creators and organizes the entries into a taxonomy.

3. Site Analysis Tool: The third tool helps us understand the structure of the website and allows us to assign metadata to the series. It also alerts us when new series appear, when series disappear, and when sites have changed radically and the analysis process must be repeated.

4. Packager: This tool performs all the functions I've described under acquisition. The software will create a METS package that can be loaded into DSpace, Fedora, Greenstone, OCLC's Digital Archive, or other METS-compliant digital repository software.

OCLC has released to its partners the alpha version of the first two tools, with the final two to be released in January. The tools are built using open source software and the final product will also be open source. However, OCLC may also provide the tools to repositories as a service with full support.

I want to take a moment to acknowledge the team at OCLC for their hard work. They have done an excellent job of building tools that support the Arizona model

Stay Tuned!

I'll conclude by noting that the development and testing of the Arizona Model is ultimately a research project. I believe that we will produce a number of very practical tools. But, those tools will be refined as we learn more. We may also learn that some things are not possible, and discover possible things we hadn't thought of.

I believe we're trying to change our practices to match the revolutionary changes in the world around us. We're taking some big steps into unknown territory. Wish us luck!